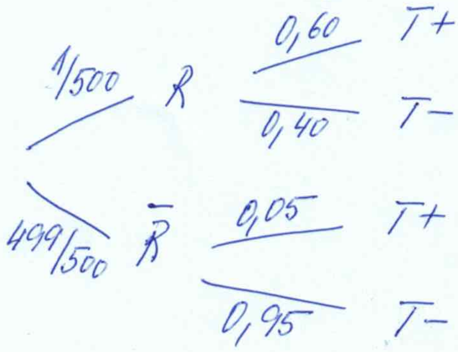
Datum: 5. 5. 2020 Osobní číslo: ML1

1. Identifikace rakoviny plic na základě rentgenového snímku je falešně pozitivní (pozitivní v případě, že pacient nemá rakovinu plic) v 5 % případů a falešně negativní (negativní v případě, že pacient má rakovinu plic) ve 40 % případů.
   a) Karlovi udělali v rámci preventivní zdravotní prohlídky rentgenový snímek hrudníku. Tento snímek se vrátil s pozitivním nálezem na rakovinu plic. Karel si zjistil, že pouze jeden z 500 zaměstnanců na srovnatelné pracovní pozici má rakovinu plic. Jaká je pravděpodobnost, že Karel má rakovinu plic? (**5b**)
   b) Stejný test, který podstoupil Karel, podstoupil i Jiří (a snímek se mu také vrátil s pozitivním nálezem na rakovinu plic). Jiří však 20 let pracoval jako horník v uhelném dolu. Jiří ví, že 15 procent jeho bývalých kolegů má rakovinu plic. Jaká je pravděpodobnost, že Jiří má rakovinu plic? (**5b**)

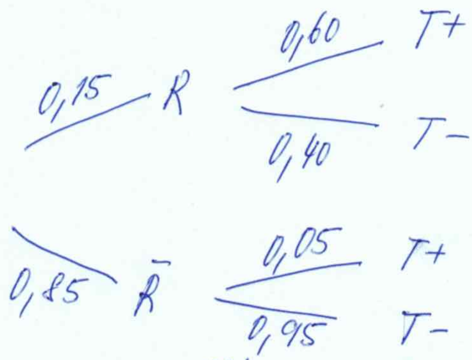
R ... pacient má rakovinu plic

$T+$ ... test vyšel pozitivní

$T-$ ... test vyšel negativní



(a)



(b)

a) $$P(R \mid T+) = \frac{P(R \cap T+)}{P(T+)} = \frac{\frac{1}{500} \cdot 0{,}60}{\frac{1}{500} \cdot 0{,}60 + \frac{499}{500} \cdot 0{,}05} \doteq$$

$$\doteq \underline{\underline{0{,}023}}$$

b) $$P(R \mid T+) = \frac{P(R \cap T+)}{P(T+)} = \frac{0{,}15 \cdot 0{,}60}{0{,}15 \cdot 0{,}60 + 0{,}85 \cdot 0{,}05} \doteq \underline{\underline{0{,}679}}$$

Datum: 5. 5. 2020 Osobní číslo: ML 1

2. V zásilce 10 výrobků je 8 kvalitních a 2 nekvalitní. Ze zásilky náhodně vybereme 2 výrobky, přičemž vybrané výrobky nevracíme zpět. Označme X počet kvalitních výrobků mezi vybranými. Určete
   a) rozdělení pravděpodobnosti náhodné veličiny X, **(2b)**
   b) distribuční funkci náhodné veličiny X a graficky ji znázorněte, **(2b)**
   c) střední hodnotu, směrodatnou odchylku a modus náhodné veličiny X. **(2b)**

   Uvažujme dále, že výrobce dostane sankci 2000 Kč za každý nevyhovující výrobek při kontrole tří náhodně vybraných výrobků z produkce.

   d) Určete očekávanou sankci výrobce, směrodatnou odchylku a modus sankcí. **(2b)**
   e) S jakou pravděpodobností bude sankce výrobce nižší než 3 000 Kč? **(2b)**

X… počet kvalitních mezi 2 vybranými

$X \sim H(N=10, M=8, n=2)$

10: KV 8, NEKV 2

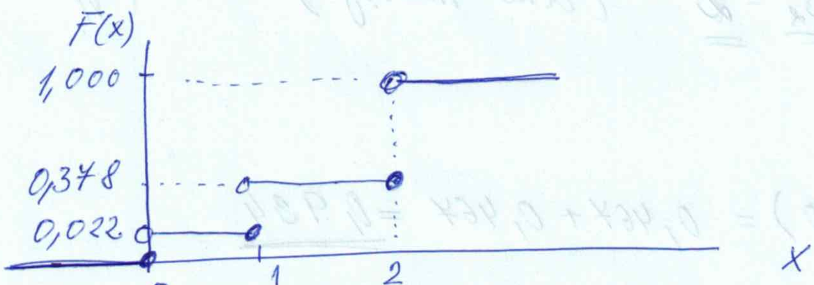
a)

x	0	1	2	Σ
$P(x)$	0,022	0,356	0,622	1,000

(výpočet v R)

b)

x	$(-\infty; 0\rangle$	$(0; 1\rangle$	$(1; 2\rangle$	$(2; \infty)$
$F(x)$	0,000	0,022	0,378	1,000



c) $\underline{\underline{E(X)}} = \sum_{(i)} x_i \cdot P(x_i) = \underline{\underline{1,600}}$

$E(X^2) = \sum_{(i)} x_i^2 \cdot P(x_i) = 2,844$

$D(X) = E(X^2) - (E(X))^2 \doteq 0,284$

$\underline{\underline{\sigma(X)}} = \sqrt{D(X)} \doteq \underline{\underline{0,533}}$

$\underline{\underline{\hat{X} = 2}}$

(výpočet v R)

d) S ... sankce (Kč)

Y ... počet nekvalitních mezi 3 kontrolovanými

$Y \sim H(N=10, M=2, n=3)$

$S = 2000\,Y$

y	0	1	2	Σ
$P(y)$	0,467	0,467	0,066	1,000

s	0	2000	4000	Σ
$P(s)$	0,467	0,467	0,066	1,000

$\underline{\underline{E(S)}} = \sum_{(i)} s_i \cdot P(s_i) \doteq \underline{\underline{1\,200 \text{ Kč}}}$

$E(S^2) = \sum_{(i)} s_i^2 \cdot P(s_i) \doteq 2\,933\,333 \text{ Kč}^2$

$D(S) = E(S^2) - (E(S))^2 = 1\,493\,333 \text{ Kč}^2$

$\underline{\underline{\sigma(S)}} = \sqrt{D(S)} \doteq \underline{\underline{1\,222 \text{ Kč}}}$

$\underline{\underline{\hat{S}_1}} = \underline{\underline{0 \text{ Kč}}}$ $\underline{\underline{\hat{S}_2}} = \underline{\underline{2000 \text{ Kč}}}$ (dva mody) (výpočet v R)

e) $P(S < 3000) = 0,467 + 0,467 = \underline{\underline{0,934}}$

Datum: 5.5. 2020 Osobní číslo: ML 1

3. Dlouhodobým pozorováním bylo zjištěno, že doba potřebná k objevení a odstranění poruchy stroje (náhodná veličina X_i) má střední hodnotu 40 minut a směrodatnou odchylku 30 minut.
   a) Určete pravděpodobnost, že celková doba potřebná k objevení a odstranění 100 poruch nepřekročí 65 hodin. (**5b**)
   b) Určete, jakou dobu si vyžádá objevení a odstranění 100 poruch, jestliže žádáme, aby tato hodnota nebyla s pravděpodobností 0,95 překročena. (**5b**)

X_i ... doba potřebná k objevení a opravení i-té poruchy (min)

$E(X_i) = 40; \quad \sigma(X_i) = 30$

X ... doba potřebná k objevení a opravení 100 poruch (min)

$X = \sum_{i=1}^{100} X_i \Rightarrow$ dle CLV $X \sim N(\mu = 100 \cdot 40; \sigma^2 = 100 \cdot 30^2)$

$X \sim N(\mu = 4000; \sigma^2 = 90000)$

a) $P(X \leq 65h) = P(X < 65 \cdot 60) = P(X < 3900) \doteq \underline{\underline{0,369}}$

b) $P(X \leq T) = 0,95 \Rightarrow T$ je 95% kvantil

$\underline{\underline{T \doteq 4490 \text{ min}}}$ (téměř 75 h)

Datum: 5. 5. 2020 Osobní číslo: ML1

4. Do podniku XYZ byla dodána zásilka velkého rozsahu (cca 50 000 ks výrobků). V náhodném výběru o rozsahu 1 000 kusů bylo zjištěno 26 nekvalitních výrobků.
   a) Určete bodový odhad podílu nekvalitních výrobků v zásilce. (**1b**)
   b) V jakém intervalu můžeme s 99% spolehlivostí očekávat podíl nekvalitních výrobků? (**4b**)
   c) Jsme oprávněni s 99% spolehlivosti tvrdit, že zásilka obsahuje více než 2 % nekvalitních výrobků? Ověřte čistým testem významnosti. (**5b**)

π ... p-st nekvalitního výrobku v zásilce

a) $\underline{\underline{\hat{\pi}}} = p = \frac{26}{1000} = \underline{\underline{0{,}026}}$ ~ 2,6 %

b) Předpoklady pro použití IO:

- $n > \frac{9}{p(1-p)} \Rightarrow \underset{1000}{n} > \frac{9}{0{,}026 \cdot 0{,}974} = 355$ ✓
- $n < 0{,}05 \cdot N \Rightarrow \underset{1000}{n} < 0{,}05 \cdot 50\,000 = 2500$ ✓

$P(0{,}01\cancel{4} < \pi < 0{,}0\cancel{42}) = 0{,}99$

Tj. 99% intervalový odhad π je $(1{,}\cancel{4}\,\%;\ \cancel{4{,}8}\ 4{,}2\,\%)$.
Cloppernův-Pearsonův

c) $H_0: \pi = 0{,}02$
$H_A: \pi > 0{,}02$
(pozor! $\alpha = 0{,}01$)

Na hl. významnosti 1 % nelze zamítnout H_0 (Cloppernův-Pearsonův test, p-hodnota $\doteq 0{,}110$).
Tj. $\hat{\pi}$ statisticky významně nepřevyšuje 2 %.

5. 4 modifikované odrůdy řepky (označeny jako A, B, C, D) byly pokusně pěstovány na několika pozemcích stejné velikosti (1 ar). Výnosy semen (v kg) z jednotlivých pozemků jsou zaznamenány v repka.xlsx (http://homel.vsb.cz/~lit40/DATA/repka.xlsx). Vhodným testem na hladině významnosti 0.05 rozhodněte, zda se střední hodnoty výnosů jednotlivých druhů řepky liší. Pokud ano, určete, které odrůdy se od ostatních liší a jak, tj. na základě post hoc analýzy seřaďte odrůdy od nejlepší po nejhorší na základě očekávaných středních výnosů. **(10b)**

Data obsahují odlehlá pozorování (identifikována metodou vnitřních hradeb. 1x skupina B, 1x skupina C). Odlehlá pozorování byla vyřazena z dalšího zpracování!

Ověření předpokladů:

- nezávislost: Data jsou nezávislá (každá hodnota měřena na jiné ex. jednotce).
- normalita: (*)

odrůda	Shapirův-Wilkův test (p-hodnota)
A	0,360
B	0,297
C	0,612
D	0,410

Na hl. v. 5% nezamítáme předpoklad normality (viz tab. (*)).

- shoda rozptylů: Na hl. v. 5% zamítáme předpoklad o shodě rozptylů (Bartlettův test, $x_{OBS} = 15,3$, $df = 3$, p-hodnota $= 0,002$).

Vlastní test – Kruskalův-Wallisův test:

$H_0: \; x^{A}_{0,5} = x^{B}_{0,5} = x^{C}_{0,5} = x^{D}_{0,5}$

$H_A: \; \neg H_0$

Na hl. v. 5% zamítáme H_0 (p-hodnota $\ll 0,001$).

mediány výnosů řepky se pro jednotlivé odrůdy st. významně liší.

Post hoc analýza (Dunnův test)

Písmenkové schéma

B (68)	a	
A (66)	a	~~b~~
C (58)		c
D (58)		c ~~d~~

St. významně nejvyšší výnosy poskytují odrůdy A a B (mezi nimi není st. v. rozdíl), st. významně nižší výnosy než A a B poskytují odrůdy C a D (mezi nimiž není st. významný rozdíl).